Local analysis of parameter covariances resulting from the calibration of an overparameterized water quality model

U. Callies, M. Scharfe
GKSS Research Center, Institute for Coastal Research
Max-Planck-Str. 1
D-21502 Geesthacht, Germany

phone: +49 (0)4152 / 87-2837
fax: +49 (0)4152 / 87-2818
e-mail: callies@gkss.de

Mechanistic water quality simulation models are important tools for supporting environmental management decisions. Possibly the most severe problem with the usage of mechanistic models is that in most cases they cannot be fully identified from data due to model overparameterization. The calibration of overparameterized models results in covariances among model parameters. The neglect of these effects may lead to a significant overestimation of model output uncertainty. We discuss principal component analysis (PCA) of the posterior parameter error covariance matrix as a tool for the identification and proper representation of parameter covariances. Our study deals with a water quality model specifically designed to support the interpretation of algae biomass observations at one single station (Weir Geesthacht) on the Elbe river in Germany.

The motivation for our modeling activity has been to test a hypothesis according to which observed negative correlations between temperature and chlorophyll *a* concentrations in summer might be an indication of growth limitation by lack of silica. A specific model concept has been implemented, according to which water bodies travelling downstream towards station Weir Geesthacht are initialised by a certain concentration of silica. Diatoms in the water bodies are assumed to cease growing and to start to decay as soon as this initial reservoir of silica has been used up. The higher growth rates are the earlier the diatom maximum occurs and the more pronounced it is. If growth rates are large enough that all available silica is assimilated already upstream of Geesthacht, further increases of growth rates (i.e. more favorable growth conditions) imply decreasing diatom populations at the end of the particle's journey at station Geesthacht.

The mechanistic water quality model we used is rather simple and incorporates aspects of an empirical approach. However, the study confirms the well known fact that even simple model structures are often not identifiable from available data. The model involves the representation of two different algae species, green algae and diatoms, both of which are known to significantly contribute to the total amount of algal biomass in the river Elbe. Only growth of diatoms depends on the availability of the nutrient silica. However, during periods with sufficiently high concentrations of silica the reactions of the two algae species to more or less favorable weather conditions are similar so that the observations of total algal biomass used in this study, are insufficient for disentangling all differences between the two species.

Our study is intended to illustrate an approach for identifying and coping with model components that are not controlled by data. The model is not parsimonious in the light of the existing data and shares the problem of overparametization with more detailed mechanistic models. The objective of the investigation reported in this paper is not to find the "true" parameter values (even if the concept of "true" parameter values was applicable) but to analyze the parameter interaction structure, which results from model calibration. Specific combinations of parameters may be much less uncertain than the individual parameters they are made up by. The opposite is also true: Some parameters may be collectively more uncertain than any of the individual parameters.

A quadratic cost function, *J*, is used to assess the deviation between model output and observations scaled by an observational error, σ_{obs}. The relevance of parameter covariances for a moderately non-linear model's fit to data can be analyzed by examining the curvature of the loss-function at its minimum. Let $\vec{x}$ denote the vector being made up by those model input parameters, x_{i}, which are to be adjusted in the process of fitting the model to data. If $\vec{m}$ denotes the vector of model outputs, m_{t}, at times *t*, the Jacobian matrix $\partial\vec{m}/\partial\vec{x}$ contains all information about local model sensitivities and provides also the basis for the estimation of parameter uncertainties and emerging model prediction errors. Assuming that model output depends linearly on the model input parameters, the posterior error covariance matrix, $\mathbf{V}_{\text{post}}$, of retrieved parameters can be obtained as the inverse of the Hessian matrix, **H**:

$$\mathbf{V}_{\text{post}}^{-1} = \mathbf{H} = \frac{\partial^2 J}{\partial \vec{x}^2} = \frac{1}{\sigma_{\text{obs}}^2} \sum_{\text{t=1}}^{\text{N}} \left(\frac{\partial m_{\text{t}}}{\partial \vec{x}} \right)^{\text{T}} \left(\frac{\partial m_{\text{t}}}{\partial \vec{x}} \right) + \alpha \mathbf{V}_{\text{prior}}^{-1}$$

The inclusion of prior knowledge with uncertainty $\mathbf{V}_{\text{prior}}$ into the definition of the cost function (using some weighting factor α) is necessary to render the inversion of the Hessian matrix possible. An alternative to using background knowledge for the derivation of the posterior parameter error covariance matrix would be to set small eigenvalues of the Hessian to zero and then to calculate the pseudoinverse of the truncated Hessian. In our opinion, however, this latter approach based on singular value decomposition and needing the specification of the level of truncation would be less transparent.

If the model of interest is linear, the posterior parameter error covariance matrix does not depend on the point in the parameter space. However, in the vicinity of the minimum the cost function should be quadratic so that the above formula provides a good approximation to the local inverse of the posterior parameter covariance matrix even in the case of a weakly non-linear model. We analyze $\mathbf{V}_{\text{post}}$ in terms of its eigenvectors and eigenvalues (PCA). PCA of the posterior parameter error covariance matrix gives a clear picture of how many degrees of freedom are really controlled by data. However, to make parameter uncertainties comparable we have first to remove all physical dimensions by proper scaling. Scaling of parameter necessarily introduces some subjectivity into the analysis. For the present study we decided to measure changes of parameter values in terms of multiples of their prior uncertainty. This is a very natural choice of units and allows for a straightforward identification of those degrees of freedom in the model (either in the original parameter space or in the space of principal components), that are identifiable from the data.

In the present study six relevant model input parameters have been selected for model calibration. PCA, however, suggests the existence of essentially two degrees of freedom in this six-dimensional parameter space that are relevant for a successful reproduction of chlorophyll *a* observations. This general picture does not change when two additional parameters are included into the calibration exercise. Principal components as artificial new input parameters do not necessarily have a physical meaning. In the present example, however, the two aggregated parameter combinations being controlled by data could be interpreted in terms of the distinction between two algae species in the model.

An analysis of model prediction errors caused by uncertainty in the space of principal components provided a better understanding of the mechanism of parameter calibration. If a model is linear, uncertainties of model parameters can be propagated independently and their effects on model output can be superimposed to each other. A main result from the analysis of model prediction uncertainty in the space of principal components has been that the two leading principal components are relevant for model output (i.e. are identifiable from the data) in distinct periods characterized by the presence and lack, respectively, of simulated silica. The corresponding loading coefficients summarize parameter changes, which all enhance chlorophyll *a* concentrations during periods with and without lack of silica, respectively. Accordingly parameter changes affecting the periods of lack of silica must have an impact on the posterior parameter covariance matrix. This contradicts the assumption of a strictly linear model and indicates the limitation of a local uncertainty analysis applied to the present example.